学校编码：10384　　分类号_____ 密级_____

学号：19020121152465　　UDC_____

硕 士 学 位 论 文

基于Dirichlet混合过程的半参双自回归模型的估计

Estimating a semiparametric double-autoregressive model with a Dirichlet process mixture

刘文英

指导教师姓名：刘 继 春 教授

专 业 名 称：概率论与数理统计

论文提交日期：2015 年 4 月

论文答辩日期：2015 年 5 月

学位授予日期：

答辩委员会主席：_____

评 阅 人：_____

2015 年 5 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密　（　），在　　年解密后适用本授权书。

2、不保密（　）

作者签名：　　日期：　　年　　月　　日

导师签名：　　日期：　　年　　月　　日

中文摘要

贝叶斯方法是当今统计学界使用相当广泛的统计手段，在很多方面硕果累累。不过非参数贝叶斯方法在时间序列分析领域内的应用还存在差距，这主要是因为在参数空间上寻找有效先验分布困难度较大。近几十年学者致力于研究先验分布的选择问题，并有所突破。本文就是基于Ferguson1973年提出的Dirichlet过程先验来研究时间序列模型，在不限定模型扰动项的方差情况下使得模型更加灵活有效。

资产收益波动率的研究一直是学术界的热点，其中最为著名的就是GARCH类模型。然而近期另一种新的异方差模型——双自回归模型受到人们的关注。Ling在2007年研究了该模型的一般平稳遍历条件，并给出了模型的极大似然估计。但该模型的很多结果局限于扰动项η_t服从正态分布的情况下，其非正态的研究尚少。本文将Dirichlet混合过程应用于双自回归模型，使得该模型能够处理时间序列数据存在异质性(如周期性、多峰分布等)的情况。同时模型中引入潜在变量，结合Gibbs抽样方法，使得抽样更有效率，且模型能够很好地对时间序列的异质性进行解释。本文应用该模型对模拟数据和实证数据进行了分析，发现该方法能很好地解释数据中潜藏的信息。由于该非参数方法对模型的假设较少，因此比通常的参数方法更加稳健。

关键词：非参数模型；Dirichlet混合过程；双自回归模型

Abstract

Bayesian method is widely used in statistics, but the nonparametric bayesian method in time series analysis is still a gap. Because it's difficult to find an effective prior distribution. In recent years, scholars have made a breakthrough in the choice of the prior distribution problem. Dirichlet process to the time series model of this article is based on Ferguson's theory. We do not attempt to qualify the mean or variance of the disturbance, which makes the model more flexible.

The research about the heterogeneity of asset returns volatility has been popular in academic world, and one of the most famous heteroscedastic model is the GARCH class model. Recently another new heteroscedastic model catches people's attention, calling double autoregressive model. Ling studied the stationary and the ergodicity of this model, then gave the maximum likelihood estimation of the model. And most results of the model must be limited to the disturbance that η_t should obey the normal distribution. The non-normal research is still little. This paper applies Dirichlet process mixtures to double autoregressive model, which allows the model to deal with time series data existing heterogeneity, such as periodic and multi-peak distribution,etc. Using Gibbs sampling method, potential variable is introduced and this model explains the heterogeneity well. In this paper, the model was applied to analyze the simulation data and the empirical data, and found that this method can better explain the hidden information in the data. The nonparametric method for model has less hypothesis, so the parameters of the method is more robust than usual.

Key words: Nonparametric model; Dirichlet process mixture; Double autoregressive model

目 录

Contents

第1章　引言

1.1　选题背景

统计学的根本目标是利用生活中观测到的真实数据去推断总体的某些性质，推断过程中常需对所研究的总体做一些假设。数理统计研究的许多方法是对总体假定一个参数模型，模型服从的具体分布在假定下是已知的，但模型中的参数未知，问题就变成估计这些未知参数或对他们进行某种假设检验。然而非参数统计方法一般对研究的总体不作具体的模型假设，只有一些定性的描述，在弱假设下对总体的一些未知特征进行统计推断。由于对总体分布的假定很少，较难依据假定由逻辑推理寻求最优推断。一般都是先由直观确定一个可行的推断方法，然后再判断方法的好坏，通常只能对各个方法进行相对比较，确定相对的优劣。

参数模型已为人们长期应用于经典统计和贝叶斯统计，它的估计是根据样本数据估计总体分布中包含的未知参数。人们根据手中的数据，分析或推断数据反映的本质规律，也就是总体的分布或数字特征。参数贝叶斯方法通常需要假设参数服从先验分布，而复杂问题服从的真实分布是难以确定的。为了计算上的方便，参数先验分布一般会采用指数族分布或者共轭分布，但这种先验假设具有较大的主观性。对于复杂数据的拟合问题，结果均难令人满意。

非参数贝叶斯方法假设参数服从一类样本空间上的宽先验分布，不需要假设参数具体的先验分布，参数的后验分布可以通过统计推断解析。非参数贝叶斯方法可以将一个复杂分布分解成无限个分布分量，并确定各分布权重。在设定拟合误差范围内，分量的个数可以自动确定，这是其他分布聚类算法不具备的优点。近几年，非参数贝叶斯统计在统计学、机器学习、生物信息、信息检索等领域的应用相当广泛。与传统的参数方法相比，非参数贝叶斯方法具有以下优点：

- 稳健性：在使用传统的参数方法时，需要对总体的分布做出假设，而这些假设并不能得到很好地满足。非参数方法则不依赖于对总体的假设，而是根据观测数据来进行推断，避免了因假设不当而产生错误。
- 适用性强：非参数方法不仅可以处理适用参数方法的问题，而且在处理定类、定序数据等问题时有良好的表现。

在非参数贝叶斯研究领域中，基于标准成分混合的未知数据分布的模型已经被学者广泛研究，例如正态分布混合、主流密度估计方法，也包括核技术[1]、非参极大似然估计[2]和基于Dirichlet混合过程(DPM,Dirichlet process Mixture)的贝叶斯方法。相对于更传统的非参估计方法，例如核技术和平滑参数估计，Ferguson[3]的方法提供了坚实的理论基础。无论是作为一种非参数密度估计方法还是作为解释变量间独立性的潜在归类识别方式，一般分布的混合作为分布建模对解决问题十分有用。贝叶斯框架内，可数的无限数目成分混合可以通过使用Dirichlet过程这样的先验分布来处理。在很多情况下，可数无限混合模型比小数目成分的混合模型更加贴近实际。

在常规参数模型对分布假设强加不合理约束的情况下，DPM在建模上的使用已变得越来越受欢迎。其中包括经验贝叶斯问题[4]、非参数回归[5]、密度估计[6]、分层模型[7]、删失数据分析[8]以及估计非标准连接函数问题[9]。

非参数贝叶斯方法综合了非参数方法和贝叶斯理论的思想，由于计算算法在应用和效率方面实质性的改善，非参贝叶斯模型近几年在实践中的应用相当广泛。例如，在分层贝叶斯模型中使用Dirichlet过程(DP)先验[10]来描述分布的不确定性，这样的DP混合(DPM)模型[11]使用各种马尔科夫链蒙特卡洛(MCMC)算法[12]来进行模型估计。基于Dirichlet过程的研究已经有非常多了，例如，纵向随机效应模型中，随机效应无论在贝叶斯模型或者经典模型中一般都假设为服从正态分布。Ken和Joseph则赋予随机效应非参数先验分布[13]，即Dirichlet过程先验。Mouchart,Rolin在研究生存模型时应用Dirichlet过程探讨删失数据和异方差性[14]，同时还研究了比例危险率模型的贝叶斯半参数形式。Basu,Chib将DPM应用于边际似然和贝叶斯因子[15]。

Hanson和Johnson对区间删失数据的加速失败时间模型引入DPM，使用MCMC方法对模型进行后验推断[16]。Lee,Berger在模型选择上用DPM作加权函数的先验分布，同时引入与选择机制相关的潜在变量[17]。Mark,Jensen年通过建立非参收益波动率分布扩展了非对称的SV模型[18]，主要手法是使用一个无限混合的正态分布，该分布是一个具有Dirichlet过程先验的未知混合。在不同领域中的应用，使用DPM模型均取得了不错的结果，且结果的分析更贴近实际。可见DPM在数据建模方面适用性佳，且应用更广。

资产收益的波动率不仅是时变的，而且在同一段时间内会出现偏高或偏低，呈现波动率聚类和尖峰厚尾现象。为了刻画这两类现象，Engle提出了自回归条件异方差(ARCH) 过程，ARCH模型已然众所周知[19]。对于时变波动率的建模，最具代表性的模型应该是GARCH类模型。最近又出现了另外一种值得考虑的模型——双自回归模型，它能很好的刻画时间序列的非平稳性，这关乎到检测一个金融市场是否是有效的。作为ARCH类模型，它可以解释时间序列中的异方差性，同时刻画出时间序列的异质性，例如尖峰后尾现象。双自回归模型(Double AR Model)较早见于Weiss,Cline的文献，主要研究了一般双自回归模型的概率结构[20]，Li,Ling则具体研究了DAR(p)模型的结构和性质[21]，他们的研究都限制扰动项的分布为正态分布。Jensen2012[18]年利用Dirichlet过程来研究SV模型，很好地刻画了收益波动率的金融特征，并且放宽了对扰动项的人为设定，使得模型更为稳健。受到Jensen的启发，本文尝试将DP过程应用于双自回归模型，作为该模型的非参数先验，令该模型的扰动项在正态分布无限混合的条件下对金融时间序列数据的特征进行刻画与还原。

1.2　研究思路

基于Dirichlet过程研究非参数问题的主要思想如下：令X是可测空间$(\mathcal{H},\mathcal{A})$上的一个随机变量，代表“数据”，它的分布完全未知，未知的概率分布也被认为是参数，取值于$(\mathcal{H},\mathcal{A})$上的所有概率测度构成的空间。Ferguson首次提出了Dirichlet过程

的形式化定义，并指出任意Dirichlet过程先验分布的后验分布仍然服从Dirichlet过程。Ferguson利用Gamma分布的逆Levy测度构造Dirichlet过程，计算复杂且不存在解析解。1994年Sethuranman[22]提出了一种便捷的Dirichlet过程构造方法，称之为截棍构造法(stick-breaking construction)，该构造法具有两个参数：强度参数α、基础分布G_0，记作$DP(\alpha, G_0)$，可独立构造服从Dirichlet过程的随机样本，其中随机原子序列由基础分布G_0抽样得到，随机概率质量则以α为参数的beta分布产生。这个定义是Dirichlet过程及其扩展模型的基础。理论上，截棍构造方法可以以任意精度逼近真实分布，但是为了计算方便，通常在允许的逼近误差范围内作有限抽样，称之为截断截棍构造方法(truncated stick-breaking construction)。相对Ferguson的表示法，Sethuraman的表示法在实际中的应用价值更高，该法更为人所理解，本文所使用的Dirichlet过程定义就是使用的后者。

本文的主要工作是在Dirichlet混合过程的框架下使用正态混合来研究数据的分布，使用了Escobar,West的方法[23]。此处的标准正态混合模型与Ferguson的类似。假设数据$y_1, \cdots, y_n$是条件独立且满足正态分布，$(y_i|\pi_i) \sim N(\mu_i, V_i)$，均值$\mu_i$，方差为$V_i$，令参数为$\theta_i = (\mu_i, V_i), i = 1, \cdots, n$。更进一步假设，$\theta_i$来自$\mathcal{R} \times \mathcal{R}^+$上的先验分布。如果对于参数的一般先验分布是部分或者全部未知的，例如Dirichlet过程，那么这个数据就来自一个正态Dirichlet混合[23]。其中很重要的特例就是正态分布中的方差是给定的$V_i = V$，该特例已经得到了广泛的研究，可以参看West[24]的研究成果，他考虑的情况是正态混合的均值μ_i来自一个未知先验，该先验就是在正态基础测度之上的Dirichlet过程[25]。与核估计结合的方法也在这些文章中有研究，这些分析包含了V_i估计的问题。Escobar考虑了相似的模型，假设$V_i = V$是已知的，不同在于Dirichlet过程基础测度的使用上有所不同。Ferguson使用Kuo研究出的蒙特卡洛方法[26]，考虑了更一般的相异未知V_i的情况。随着理论研究基础的坚实，计算技术也随之发展。Escobar,West发展了Monte Carlo分析，拓展了Escobar的研究。MacEachern则解决了应用Gibbs抽样处理基础测度非共轭DPM 模型的公开问题[7]。Walker,Damien应用了一个相对不同的辅助变量方法

估计某些DPM模型[27]。Neal则结合了上述方法，使用了基于辅助变量的Gibbs抽样，更有效率，并且规避了可能存在的近似误差。本文借鉴了Neal的抽样方法，建立固定均值$\mu_i = 0$，方差V_i来自一个未知先验的双自回归模型，该先验是在逆伽马基础测度上的Dirichlet过程，后文将具体给出建模和抽样的过程。

1.3　文章结构

综合上述，虽然非参数方法和贝叶斯推断理论各自在时间序列分析中都有丰硕的研究成果，但是非参数贝叶斯方法在这方面的应用却不多。本文立足于非参数贝叶斯方法针对双自回归模型讨论了其在时间序列分析中的应用，对于时间序列分析的发展有一定的意义。根据以上的布局，先简要介绍下文章的主要脉络：

第一章为引言，介绍了课题的选题背景以及与该课题相关的一些研究情况，引出本文所要讨论的问题。

第二章简要介绍了非参数贝叶斯方法的发展情况，以及Dirichlet过程的基本框架。

第三章引出本文讨论的基于Dirichlet过程的双自回归模型的构造、先验分布的选取以及参数的抽样算法等问题。

第四章是通过模拟数据的求解来验证模型是否正确以及是否能反应所研究的问题，并选取实际数据，针对现实问题使用本文提出的模型进行求解。

第五章对全文主要工作进行总结概括。

第2章　非参数贝叶斯方法和Dirichlet过程

2.1　贝叶斯理论

现代贝叶斯理论的产生可追溯到20世纪30年代初，Ramsey(1931)在《数学基础》一书中首次提出将主观置信度视为概率的一种解释，使其成为频率解释的一个补充。直到50年代，贝叶斯这个术语才正式为统计学界所使用，拥有了一批忠实的支持者。贝叶斯定理是整个贝叶斯理论的基石，它是关于随机事件A和B的条件概率和边缘概率之间关系的一则定理。

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2-1}$$

贝叶斯推断便是将上述等式实例化的一个过程。假设我们现在关心的是某个未知参数的分布，根据贝叶斯学派的观点，在推断参数分布的过程中除了使用样本所提供的信息外，还必须规定一个先验分布。先验分布可以是主观的，它包含了我们对参数的事前认识。在有了先验分布以及样本之后，我们就可以通过贝叶斯定理来确定参数的后验分布：

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}. \tag{2-2}$$

其中$P(X|\theta)$称为似然函数。$P(X)$是与θ无关的项，可通过积分$P(X) = \int_{\widetilde{\theta}} P(X|\widetilde{\theta})P(\widetilde{\theta})$来计算。但随着贝叶斯理论的发展，先验分布的形式越来越复杂，很多情况下要解析的计算出$P(X)$的形式变得异常困难，抽样算法便由此产生。

2.2　非参数贝叶斯理论

随着贝叶斯理论的不断发展，统计学家们发现在诸多情况下有必要减少对总体分布的限制性假设。在利用一些常见分布来研究总体情况时要加入一些强制性的假设，

比如说，在使用正态分布或者t分布来表示总体分布时，认为总体是单峰的。做出这样的假设后，就不能处理多峰的情况，所以我们必须要寻找参数更为复杂的总体分布。例如，在贝叶斯稳健性研究中就有人提出使用指数族作为正态分布族的推广。但此类分布无法处理非对称的情况。另外一种可能的解决方法是采用有限混合模型，这种方法在出现之后便由于其在模拟逼近基础上易于进行推断这一优点而受到广泛的关注[28]。

非参数贝叶斯推断就是基于上述第二种方法，但通常情况下所包含的参数是无限个，是建立在函数空间上的概率模型。和传统的贝叶斯推断相比，非参数贝叶斯推断放宽了对模型观测值以及未知参数概率模型的假设，因而具有更高的灵活性。Muller对非参数贝叶斯的方法进行了系统的介绍[29]。

2.3　Dirichlet过程理论

Dirichlet过程最早是由Ferguson[10]提出的，他提出了一种在贝叶斯框架下进行的非参数估计方法，即在无限维参数空间中定义先验分布，这个先验分布就是Dirichlet过程。Dirichlet过程在非参数学习中应用广泛，主要被作为非参数问题研究中的先验分布。目前，有很多应用贝叶斯方法基于Dirichlet过程研究非参数问题的相关研究，例如分层Dirichlet过程，Dirichlet混合过程等等。

Dirichlet过程是基于Dirichlet分布而生成的，且其性质与Dirichlet分布的性质类似，因此首先来介绍下Dirichlet分布。

2.3.1　Dirichlet分布

假设随机变量$X = (x_1, \ldots, x_k)$满足参数为$(\alpha_1, \ldots, \alpha_k)$的Dirichlet分布，记为$X \sim$